Healthy Bodies, Bright Minds:
"The impact of Healthcare Access on Children's Academic Performance"
June 1, 2024

Executive summary

Problem and data selection

We are interested in exploring the effect of healthcare access on children's academic performance. Research shows that school absences have negative impact on grades and student's academic achievement. For that reason, we will use absenteeism as a metric of student educational outcome. We use the dataset from the National Survey of Children's Health (NSCH) dataset, and we extract from this dataset two sets of variables:

- Predicting variables:
  - "access to healthcare" features (e.g. children's current healthcare coverage, how often the child is allowed to see providers, etc),
  - health-related features (e.g. depression, children's general health),
- Target variable:
  - "days missed in school". We convert days_missed into a categorical variable: 0 means 0-6 days missed, 1 means 7+ days missed.

Feature selection 1.0

Our dataset is a high-dimensional dataset with 29433 rows and 448 columns (447 features, 1 target). Feature selection is a crucial step for our model as it reduces overfitting, improves accuracy, reduces computational costs, and aids interpretability. We use three different methods for feature selection:

- **Handpick**: we parse through the 447 features in the NSCH dataset, picking any related to health and healthcare access,
- **Correlation analysis** (supervised filter method): we compute the linear correlation between each feature and the number of days missed, keeping features with high correlation,
- **Histogram analysis** (supervised filter method): for each feature, we measured the change in histogram shape among children with low and with high absenteeism, keeping features with sufficiently different histograms.

Through these methods, we refined our dataset to 88 features.

Model selection

We trained 5 models to predict whether children are at risk for high absenteeism using the 88 features we identified in the previous section: a logistic regression model, a random forest classifier, a support vector classifier, a k-nearest neighbor classifier, and a stratified dummy classifier to use as a baseline model. Of these, logistic regression and random forest perform similarly and both outperform the other three classifiers. We chose to use a logistic regression model since it's the more interpretable model of the two.

Feature selection 2.0

We refined our list of features by identifying groups of collinear variables and removing all but one feature from each group. Through this method, we reduced our feature count by 14, giving us 74 remaining features. We then used recursive feature elimination to systematically remove features until the model performance began to suffer. This final method of feature selection produced a list of ten total features

Model analysis

By measuring feature importance as the magnitude of model coefficients, we found that poor health was strongly related to absenteeism Specifically, we found that a higher number of missed days was predicted by poorer general health and more time spent in the hospital, as well as the presence of depression, chronic physical pain, and digestive problems. Additionally, children who reported having problems at school, needed healthcare-related decisions made on their behalf, or experienced health problems for which their family needed to cut work hours were also found to be more likely to miss school. While our goal was to determine the influence of healthcare access on educational outcomes, our remaining features after selection provide little insight into the role healthcare access plays in the same.

Conclusion and future directions

Our results suggest that while features related to health highly affect absenteeism, features related to healthcare are not the strongest predictors of child absenteeism. However, it is possible and perhaps likely that the relationship between access to health care and absenteeism was drowned out by the more potent predictors of missed days, such as the general health of the child. A future study could control for predictors which are more related to access to healthcare. Likewise, it is possible that absenteeism is a poor metric for education outcomes; future work could try other metrics, such as grades or scores on standardized tests. Finally, our data comes from 2019, prior to the COVID-19 pandemic. It would be interesting to see if there was a more clear relationship between health care access and absenteeism in more modern, post-COVID data.